RAS Subsystems: How Will They Support Next Generation Platforms?

The National High Performance Computing Workshop on Resilience

Thursday August 13th

James H. Laros III
Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Reliability Availability and Serviceability (RAS) Subsystems

- In this context
  - Systems Software used to monitor and control the platform
- To support of Application Resilience
    - Needs to be far more
- Current “true” RAS subsystem examples
  - Cray XT3/4/5 RAS subsystem
  - Blue Gene RAS subsystem
- Risk Areas
  - Hardware
  - Software

Three Laws of ~~Robotics~~ RAS Subsystems

(Somewhat Anthropomorphized) – Isaac Asimov

1. **The RAS subsystem shall not injure the platform it serves or interfere with the main purpose of that platform.**

…..ok well maybe there is only one.

- **Out of Band (OOB) becomes more important.**
- **How do we monitor and control without affecting, or *minimally affecting*, the underlying platform?**
  - **Pretty simple if we don't do much**
  - **Gets harder as we try to satisfy resilience requirements**
- **To support Resilience we need to do MUCH more**

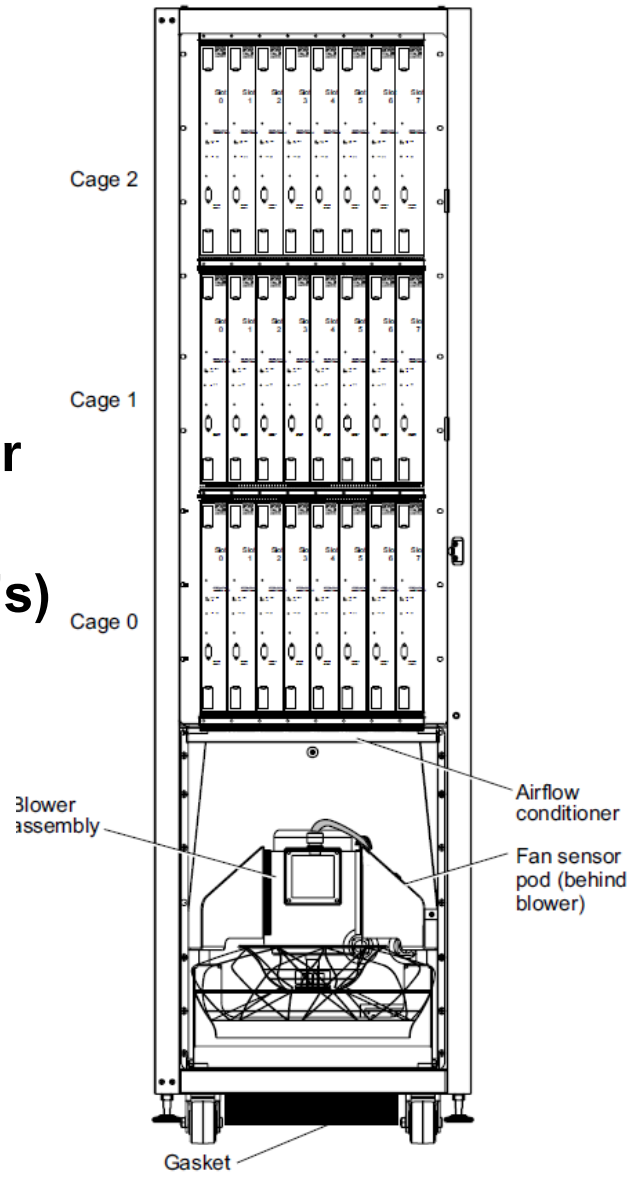
Sandia National Laboratories

RAS Subsystem is already a System

- For example
  - Red Storm (Cray XT4)

  135 (compute cabinets) * 3 (cages per cabinet) * 8 (slots per cage) * 1 (L0 per board) = 3240 L0's
- Additionally, one L1 per cabinet (135 L1's)
- One top-level System Management Workstation (SMW)
- Equivalent to a 3240+ node (disk-less) cluster
  - not counting the L1's

** Jaguar: at ≈ 200 cabinets, 4800 L0's*



Sandia National Laboratories

Projecting into the Future

- 1 Peta-Flops delivered soon-*ish*
  - Pretty much the same numbers
- 10 Peta-Flops??

  ≈500 (compute cabinets) * 3 (cages per cabinet) * 8 (Slots per cage) * 1 (L0 per board) = 12000
- Red Storm currently has 12960 compute nodes.
- Will we need a RAS sub-system for the RAS sub-system?
- Will hierarchical schemes break down?
- Failure rates have significant implications at these numbers!!

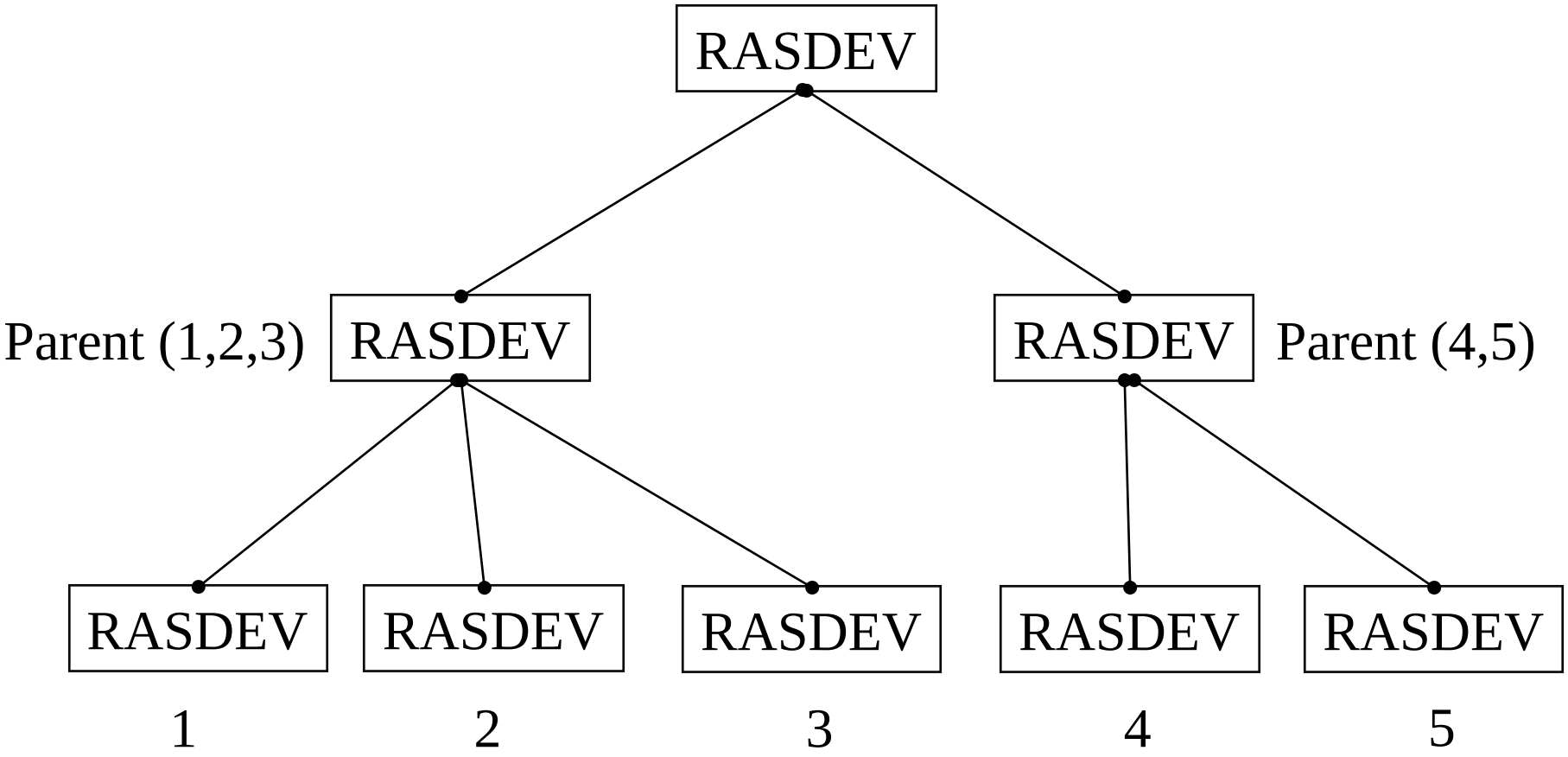
Sandia National Laboratories

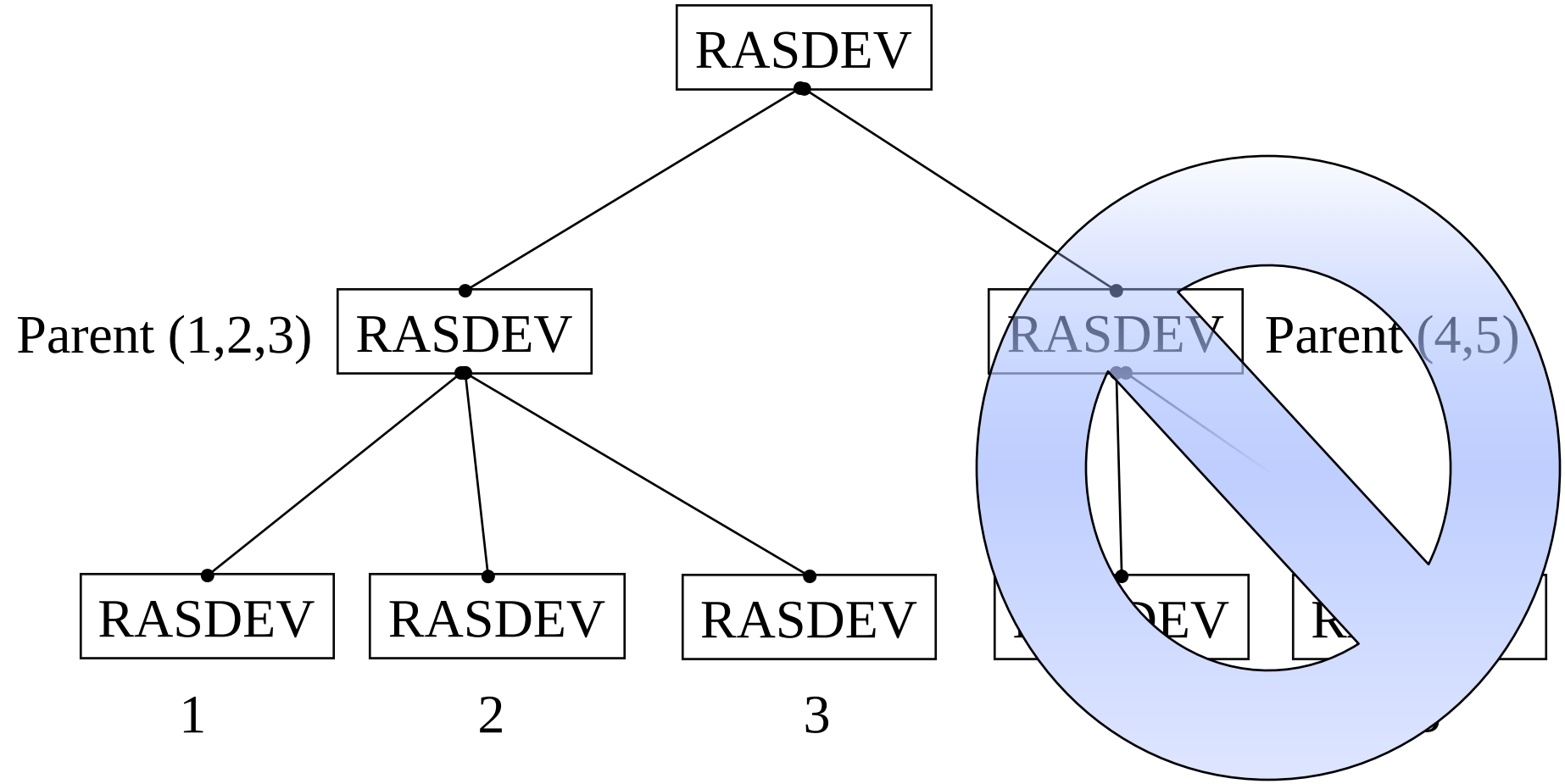
RAS for Resilience

- Resilience research *ASS/U/MES* capable RAS subsystem
  - We do, what choice do we have?
- Reality of a RAS subsystem
  - What we THINK it provides
  - What it DOES provide
  - What we NEED it to provide
- Unfortunately these tend to be very different things….
  - In addition, differ depending on platform!
- We must close this gap.

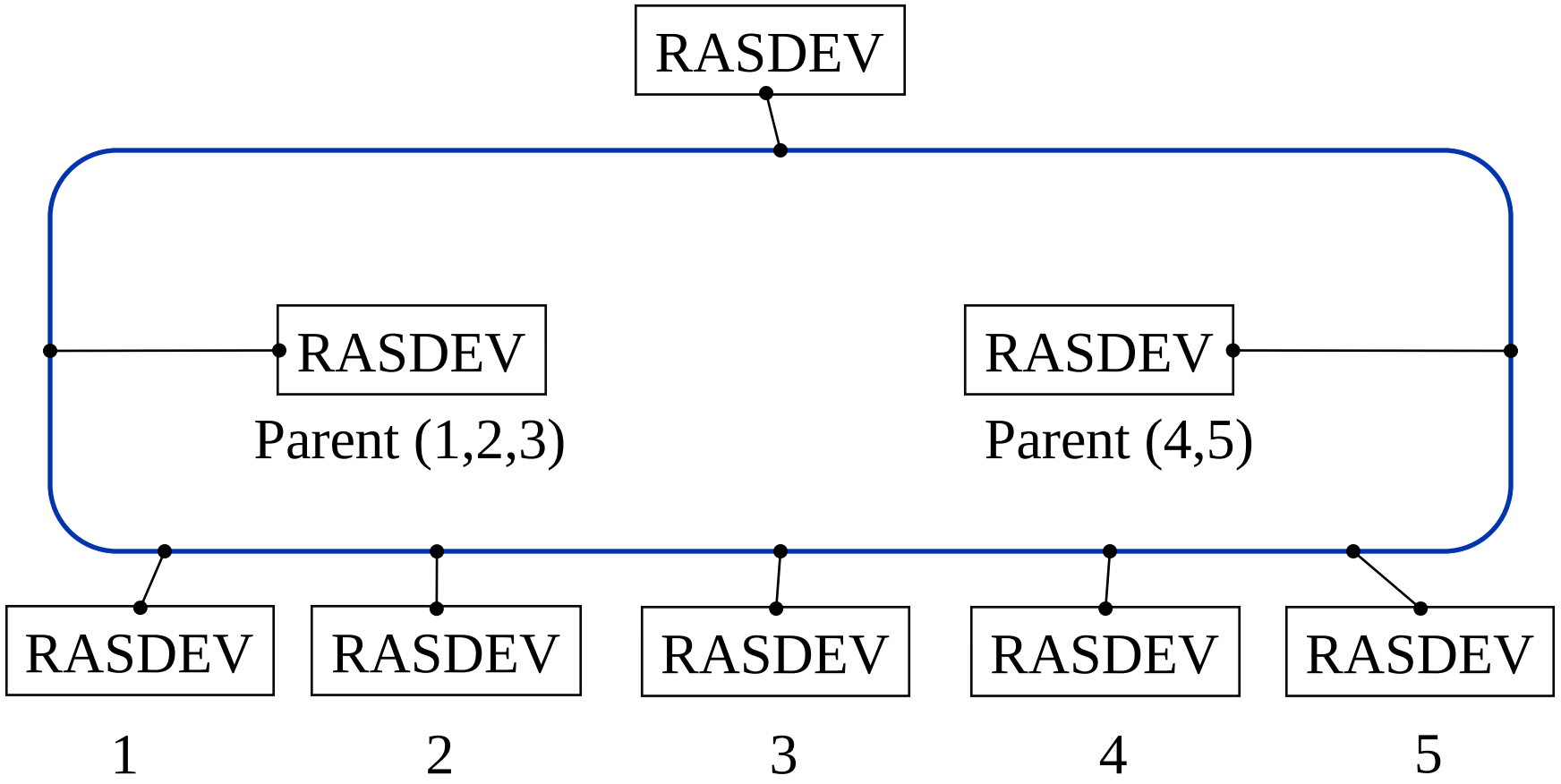
Mitigating challenges

- Configure hardware differently
  - For example, is one RAS node per board over-kill?
  - Maybe not as requirements increase to support Resilience
- Overlay Networks and other distributed systems concepts to deal with failures?
  - Dynamically re-organize network hierarchy
  - Dynamic role assumption
- Light Weight RAS message protocols?
- More intelligent RAS software?
  - Keep uninteresting things from propagating
    - What is uninteresting?

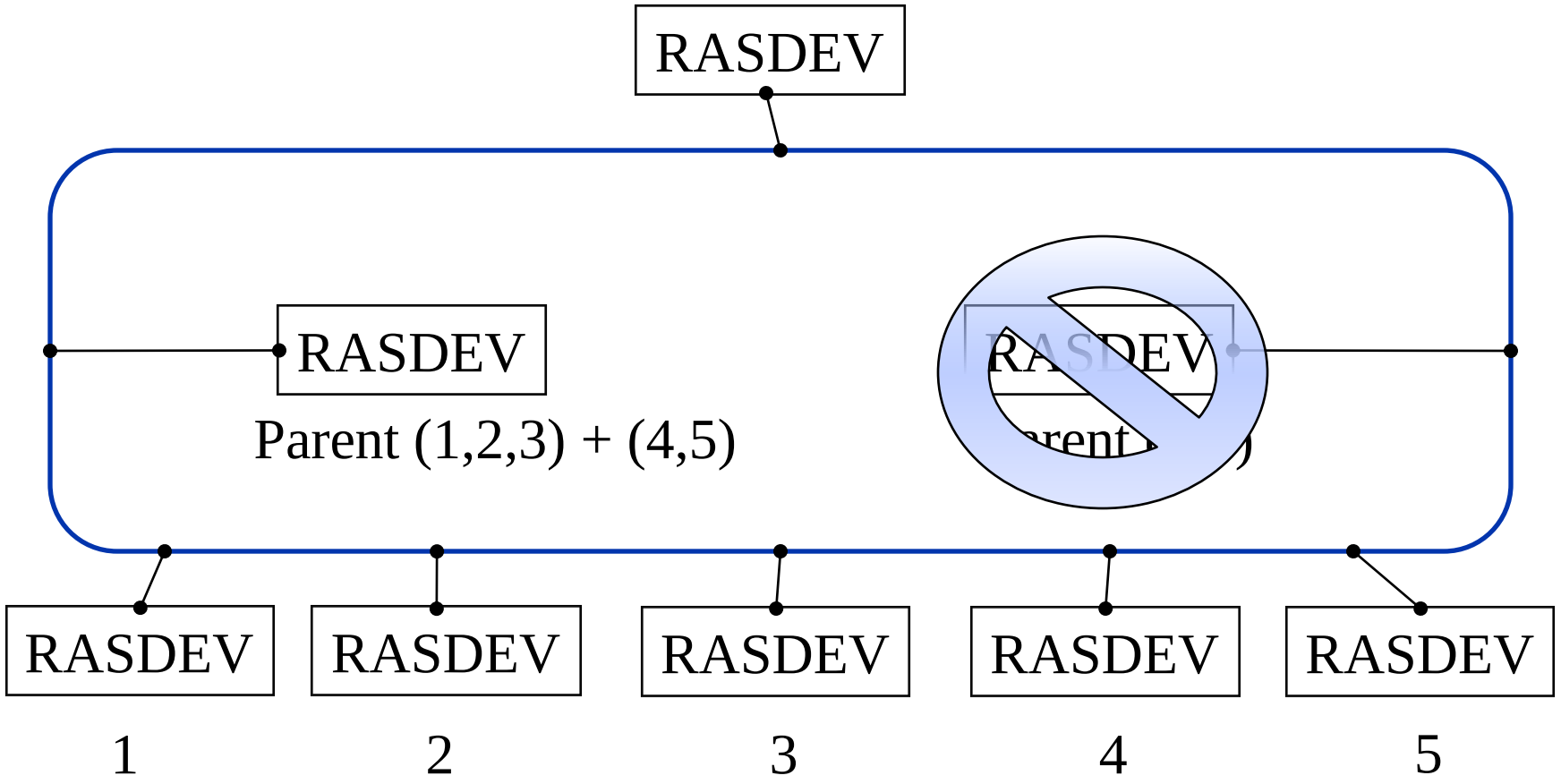




Sandia National Laboratories



Sandia National Laboratories



Sandia National Laboratories

What Does Resilience Require?

- In a word, INFORMATION
  - System Centric
    - Node characteristics
    - Physical and Logical locality of node
    - Other component information
  - Status/State type information
    - Syslog-like data
    - Sensor data
      - Huge number and numerous types
    - Component states
      - Hardware AND Software components
    - Job layout
- Once we have it, we need to GET it!
- Numerous stakeholders

Areas we can IMPACT

- RAS API
  - Standard way to interface with RAS subsystem
    - Does not dictate underlying sub-system!
      - More likely to get vendor cooperation.
    - Subscription based, Query based, etc.
    - How we make sense out of INFORMATION
  - Community researching resilience best equipped to define needs (from their perspective)
    - Other stakeholders must be involved
- Standardized Backend
  - With a complete API likely not necessary
  - Could be an area of commonality between vendors

Areas we can IMPACT
(continued)

- **RAS Communication Protocols**
  - **Possible wide applicability**
  - **University interest**
  - **At scale contributions**
- **Common RAS foundation**
  - **System Description Language**
  - **Promotes a Systems View of platform**

Conclusions/Questions

- We can have impact in this area
  - We have had some already
- Value in collaboration, and developing a standard
  - Vendors hear the same requirement from everyone

Sandia National Laboratories